3E

УДК 004.89:004.93

Т.В. Ермоленко, А.С. Гайдамака
Институт проблем искусственного интеллекта
МОН Украины и НАН Украины, г. Донецк
Украина, 83048, г. Донецк, ул. Артема, 118 б

Синтаксическая модель предложения русского языка на основе предикатных структур

T.V. Yermolenko, A.S. Gajdamaka
Institute of Artificial Intelligence
MES of Ukraine and MAS of Ukraine, c. Donetsk
Ukraine, 83048, c. Donetsk, Artema st., 118 b

Syntactic Model of Russian Sentence Based on Predicate-Argument Structure

Т.В. Єрмоленко, А.С. Гайдамака
Інститут проблем штучного інтелекту
МОН України і НАН України, м. Донецьк
Україна, 83048, м. Донецьк, вул. Артема 118 б

Синтаксична модель речення російської мови на основі предикатних структур

В статье сделан аналитический обзор существующих способов синтаксического представления предложений, предложена модель семантико-синтаксического представления предложения в виде предикатной структуры. Модель в таком виде позволит осуществлять дальнейший семантический и прагматический анализ ЕЯ-текста. Авторами разработан метод получения предикатной структуры предложения русского языка, использующий синтаксические шаблоны и словарь валентности предикатов.
Ключевые слова: лингвистический анализ ЕЯ-текстов, дерево синтаксического разбора, предикат, валентность предиката, семантическая классификация предикатов.

In the article is a review of existing syntactic representation of a sentences, a model of semantic and syntactic representation in the form of predicate-argument structure is proposed. The model in this form will allow for further semantic and pragmatic analysis of NL-texts. The authors developed a method for producing a predicate structure of the Russian sentence, which uses syntactic patterns and valency dictionary.
Key words: linguistic analysis of NL-texts, parse tree, the predicate, the valence of the predicate, the semantic classification of the predicate.

У статті зроблено аналітичний огляд існуючих способів синтаксичного представлення речень, запропоновано модель семантико-синтаксичного представлення речення у вигляді предикатної структури. Модель в такому вигляді дозволить здійснювати подальший семантичний і прагматичний аналіз ПМ-тексту. Авторами розроблено метод формування предикатної структури речення російської мови, який використовує синтаксичні шаблони і словник валентності предикатів.
Ключові слова: лінгвістичний аналіз ПМ-текстів, дерево синтаксичного розбору, предикат, валентність предиката, семантична класифікація предикатів.

Введение

Задачи автоматической обработки текстов (АОТ) возникли практически сразу после появления вычислительной техники. Так, развитие хранилищ данных делает

актуальными задачи поиска и извлечения информации, формирования корректно построенных текстовых документов. Бурное развитие Internet повлекло за собой создание и накопление огромных объемов текстовой информации, что требует создания средств полнотекстового поиска, автоматической классификации и реферирования текстов, автоматизированного машинного перевода. Таким образом, область применения систем анализа естественно-языковых (ЕЯ) текстов достаточно разнообразна, а в виду большого роста объемов текстовой информации и сложной структурированности ЕЯ-текстов, анализ текстов представляет собой очень актуальную проблему, особенно в последние годы, когда наметилась тенденция к информатизации общества.

Стремительное увеличение вычислительных мощностей сделало возможным применение трудоёмких лингвистических алгоритмов на больших объемах данных. Но несмотря на то, что в области формализации естественных языков и систем АОТ, в частности, задействовано большое количество людей и мощностей, работающих в самых разных направлениях, результаты пока довольно скудны, так как ни одна из существующих моделей не может перекрыть структуру языка в целом, а объёмы данных, с которыми имеет дело лингвистика, очень большие.

Независимо от того, на каком языке написан исходный текст, его полный лингвистический анализ проходит одни и те же стадии: графематический, морфологический, синтаксический и семантический. В результате формируются модели текста, адекватно отражающие его словообразовательные, грамматические и смысловые конструкции.

Графематический анализ – достаточно простой компонент, выполняющий первые предварительные действия над текстом. Можно выделить следующие основные функции графематического анализа [1], [2]: разбиение текста на графемы, абзацы и предложения; определение границ предложений; различение слов и служебных графем (например, знаков пунктуации); определение регистра слов; извлечение лексических конструкций (несловарных единиц, имеющих регулярную структуру: номер телефона, дата, инициалы, сокращения и т.п.); распознавание собственных имен; распознавание подписей к рисункам и таблицам; распознавание формул (математических и химических).

Корректная работа графематического анализатора невозможна без словарей фамилий, имен, отчеств, географических и административных названий, общепринятых сокращений, условных обозначений и аббревиатур, а также набора стоп-слов и шаблонов, указывающих на возможность принадлежности прилегающих слов к словарю географических и административных названий.

Графематические дескрипторы, характеризующие каждое слово входного текста, создают формальное его описание на уровне графематики, которое уже подвергается автоматизированной обработке в терминах лингвистических теорий.

Морфологический анализ – давно и хорошо отработанная лингвистическая процедура, реализованная во множестве разнообразных исследовательских и коммерческих проектов. В результате анализа для каждой словоформы текста определяется ее морфологическая информация (МИ) и осуществляется лемматизация – приведение текстовых форм слова к словарным (начальным) [2-4].

Главной проблемой является омонимичность словоформ. Например, у словоформы «стекла» два варианта морфологической интерпретации: стекло – существительное, стекать – глагол. Поэтому программы работают с целым набором возможных морфологических интерпретаций, постепенно выделяя наиболее вероятные на следующих этапах анализа.

Следующий этап обработки – *синтаксический анализ*. Его задача состоит в том, чтобы, используя МИ о словоформах, построить синтаксическую структуру каждого предложения входного текста [5].

Построение достоверных синтаксических структур всех подряд предложений текста – очень важная и нужная ступень в автоматическом понимании текста, но получить хорошие результаты синтаксического анализа для всех предложений ЕЯ-текста оказывается практически невыполнимой или безмерно сложной задачей, поскольку формальные математические модели и их программные динамические реализации не способны охватить всю сложность и многообразие языковой системы, особенно для языков с относительно свободным порядком слов, каким являются русский. В связи с присутствием в русском языке большого количества синтаксически омонимичных конструкций, наличием тесной связи между семантикой и синтаксисом, процедура автоматизированного синтаксического анализа текста является трудоемкой. Сложность алгоритма увеличивается экспоненциально при увеличении количества слов в предложении и числа используемых правил. Применение формализма для структурирования ЕЯ-предложения может привести к потере правильного синтаксического представления или комбинаторному взрыву, когда из-за морфологической и синтаксической омонимии программа оказывается не в состоянии просчитать все возможные варианты структур.

В задачу семантического анализа входит выделение смысла входного текста и выражения этого смысла на внутреннем языке системы. Выходной структурой является семантическая сеть. Одним из основных параметров анализа текста является понимание смысла входного предложения, включающее в себя описание сущностей входного текста, определение их свойств и отношений между ними. Отнесение подобных вопросов только лишь к сфере семантики неправомочно – они должны решаться на уровне синтаксической модели, так как проявляются на уровне общей схемы, не зависящей от смысла высказываний, поэтому морфолого-синтаксические признаки и структуры привлекаются в качестве правил локального контекстного разбора, задачей которого является заполнение слотов семантической сети. Таким образом, семантический анализ текста базируется на результатах синтаксического анализа, получая на входе уже не набор слов, разбитых на предложения, а набор графов, отражающих синтаксическую структуру каждого предложения. Поэтому выбор используемой синтаксической модели крайне важен для проведения качественного семантического анализа.

В данной работе предложен подход к построению синтаксической модели предложений русского языка в виде предикатной структуры.

Цель данной работы – разработка синтаксической модели предложения русского языка, позволяющей рассматривать предложение как структурированную форму сообщения, которая выражает смысл предложения. Модель в таком виде позволит осуществлять дальнейший семантический и прагматический анализ ЕЯ-текста.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Сделать обзор синтаксических моделей представления ЕЯ-предложения, обосновать выбор семантико-синтаксического представления предложения в виде предикатной структуры.

2. Разработать метод получения предикатной структуры предложения русского языка, использующий синтаксические шаблоны и словарь валентности предикатов.

Модели представления синтаксической структуры предложения

Модель синтаксической структуры предложения в значительной степени передает концепцию разработчиков лингвистических процессоров относительно синтаксического уровня анализа: какая именно информация об элементах предложения и их

взаимосвязях должна выявляться в процессе анализа, присутствовать в его результатах и какие формы представления ей адекватны. Наиболее общим для разработчиков синтаксических анализаторов является подход к получению синтаксического строения предложения с помощью некоторого частично упорядоченного множества бинарных связей между элементами. Представления о бинарных синтаксических связях используются в двух известных моделях синтаксической структуры: графах зависимостей и графах непосредственных составляющих (НС). В настоящее время эти две формы представления синтаксической структуры остаются основными, они используются в чистом виде или в смешанных формах, сочетающих в себе свойства обоих графов [5].

Графы зависимостей – способ синтаксического представления предложения как линейно упорядоченного множества элементов (словоформ), на котором можно задать ориентированное дерево (узлы – элементы множества). Каждая дуга, связывающая пару узлов, интерпретируется как подчинительная связь между двумя элементами, направление которой соответствует направлению данной дуги. Множество всех узлов дерева, прямо или косвенно зависящих от какого-либо узла, включая сам этот узел, составляет группу зависимости этого узла.

Такой способ представления синтаксических структур имеет определенные недостатки: жесткое требование рассматривать каждое формально выделенное вхождение слова в качестве отдельного элемента предложения; все без исключения связи между словоформами трактуются как подчинительные.

НС-структура – множество отрезков предложения, называемых составляющими, которое удовлетворяют следующим условиям:

– в качестве элементов множества отрезков предложения присутствуют само предложение и все его отдельные словоформы;

– в одну составляющую объединяются отрезки непосредственно синтаксически связанные между собой;

– любые две составляющие либо не пересекаются, либо одна из них содержится в другой.

С помощью НС-структур предложение анализируется как двусоставная конструкция, включающая две НС – именную и глагольную группу. Дополнение может квалифицироваться как узел, который подчинён глагольной группе. НС-структуры дают возможность выделить в предложении не только отдельные слова, но и некоторые словокомплексы, функционирующие как единое целое (например, сложное сказуемое), а также более естественно описать конструкции с неподчинительными отношениями,

К недостаткам НС-структур относятся неоднозначность трактовки силы связи между элементами словосочетаний, что приводит к неоднозначным НС-структурам (например, [[*чудовищного роста*] *смертности*] или [*чудовищного* [*роста смертности*]]), а также тот факт, что НС-структуры не вводят никакой иерархии среди составляющих одного уровня.

Общим недостатком рассмотренных моделей является то, что члены предложения определяются на основе формальных признаков: не по отношению к их возможному или реальному семантическому содержанию, а по отношению к тому месту, которое они занимают в дереве порождения предложения.

Предлагаемый подход к формированию синтаксических моделей использует предикативность – одну из важнейших характеристик простого предложения. Ни одна теория или концепция синтаксической организации предложения не обходит стороной свойство предикативности. Глагол является определяющей частью языка, предложения

без глагола или без предикативного слова не существует. Предикат – центральная синтаксема в семантическом простом элементарном предложении, формирующая его семантико-синтаксическую структуру. Предикативно связанные грамматические субъект и предикат квалифицируются как главные члены предложения, поскольку они формируют его конструктивный минимум. Более того, предикатная модель наилучшим образом отражает смысл предложения, так как в предикатах указывается не только аргументная структура и количество актантов, но и их семантическое содержание.

Предикатная модель синтаксической структуры предложения

На синтаксическом уровне предикат – это ядерная структура, которая включает в свой состав *n* актантов. Само ядро – это глагольная конструкция, а актанты объединяются с ядром системой отношений. Узлами в этой конструкции являются имена (существительное, местоимение, числительное) в их атрибутивной форме. Синтаксические отношения реализуются определенным образом, а их количество может достигать не более 7, связано это с тем, что семь – предел возможности человека одновременно воспринимать разные характеристики одной ситуации или объекта.

Предикатную модель простого предложения принимаем в следующей интерпретации (рис. 1).

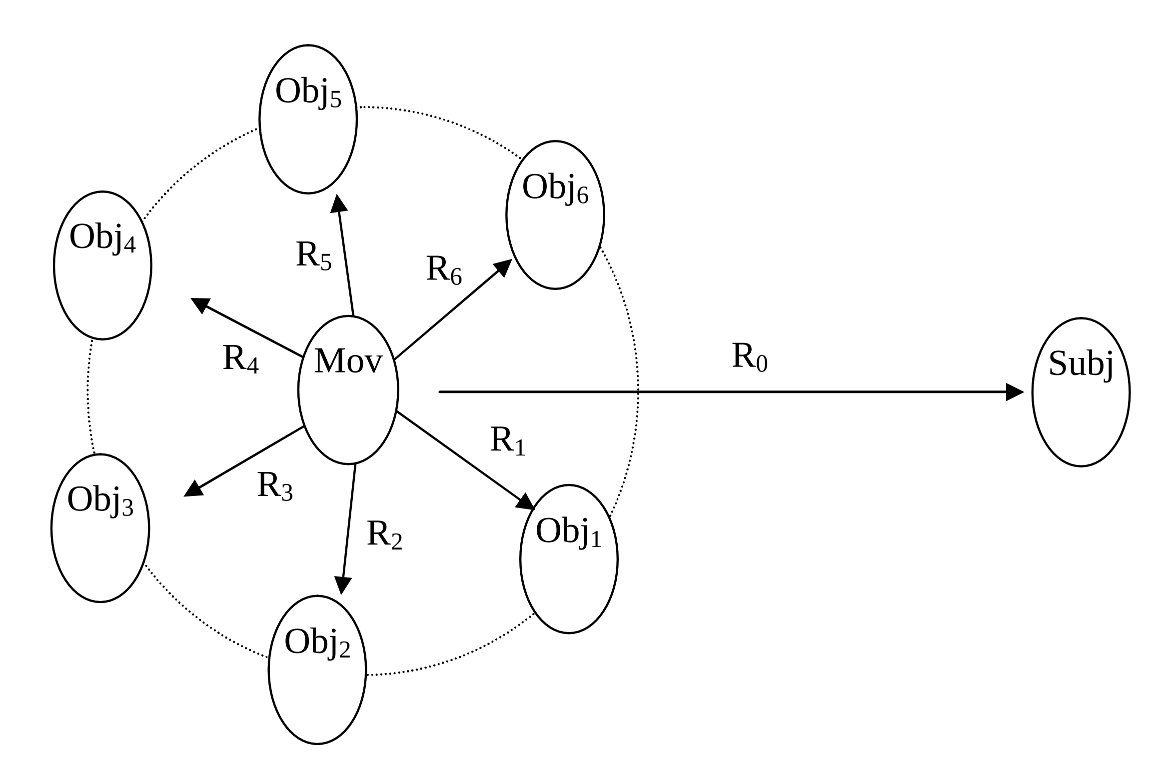


Рисунок 1 – Предикатная модель простого предложения: Mov – предикат, Subj – субъект, Obj_i – актанты предиката, R_i – отношения предиката, R_0 – отношение «быть субъектом»

Следует учитывать, что объекты, субъект и предикат человек всегда воспринимает как некоторую целостность, которая всегда реализуется через совокупность своих признаков – атрибутов, которые, например, помечают цвет, материал, отдельные стороны динамических ситуаций. Язык имеет средства для их описания (табл. 1).

Таблица 1 – Языковые средства атрибутивного описания элементов предикатной модели

Элемент модели	Обозначение	Часть речи
Объект	*Obj*	Существительное, субстантивированное прилагательное и местоимение
Субъект	*Subj*	
Предикат	*Mov*	Глагол, краткая форма прилагательного/причастия, наречия-предикативы
Признак объекта	*Attr*(*Obj*)	Прилагательное
Признак действия	*Attr*(*Mov*)	Наречие
Мера признака	*Attr*(*Attr*)	Наречие

Атрибутивный уровень формирования описаний объектов/субъектов реализуется с помощью использования хорошо известной схемы связи, которая определяется как согласование. В этом случае необходимо, чтобы сочетаемые элементы имели одинаковую МИ. Сочетаемыми элементами в этом случае будут имена.

Помимо того, что каждый объект, субъект, предикат определяются, в общем случае, на множестве своих признаков, эти элементы могу иметь зависимые слова, связанные с ними подчинительной связью типа управления и примыкания. Для общей схемы описания объекта/субъекта и предиката введем понятие звезды.

Под звездой понимается граф-звезда, узлами которого являются слова предложения, в одной доле находится главное слово, в другой доле – множество зависимых слов, отстоящих от главного на одну связь. Связи направлены от главного слова к зависимым и могут быть нескольких типов: атрибутивная (согласование), управление, примыкание.

Связи предикатной структуры имеют иерархическую зависимость, в которой четко прослеживаются три группы отношений:

1) отношение R_0, как центральное отношение двухсоставной предикатной конструкции;

2) отношения R_i предиката Mov;

3) синтагматические отношения – отношения связей внутри звезды, включая отношение атрибутивного уровня описания составляющих предложения.

Следовательно, в построенная таким образом модель позволяет полностью выявлять оба типа синтаксических отношений – предикативное и синтагматическое. Первое выражает зависимость между синтаксическими объектами через понятие, означающее действие, второе – сочетание двух синтаксических объектов, обнаруживает формальные и смысловые связи слов.

В разработанной нами предикатной модели согласно описанной выше структуре (рис. 1) содержится семь слотов, соответствующих валентным гнездам предиката. Причем номер валентности определяет ее тип, семантику и морфологическое выражение (табл. 2). Таким образом, актанты выступают в качестве семантических падежей и интерпретируются как «роли» в отношениях действия и состояния, которые выражаются предикатом.

Таблица 2 – Тип, семантика и морфологические характеристики валентных гнезд

Номер валентного гнезда	Наличие предлога	Падеж актанта	Семантический падеж
0	–	Именит.	Субъект
1	–	Винит.	Объект
2	–	Дательный	Адресат
3	–	Творит.	Инструмент
4	+	Родит.-предл.	Начальный локатив
5	+	Родит.-предл.	Конечный локатив
6	+	Родит.-предл.	Средний локатив

Немаловажную роль при формировании предикатной структуры играет семантическая классификация предикатов. В [6] аргументировано доказано, что между синтаксической формой и содержанием существует тесная связь даже на уровне классификации. Таким образом, каждому семантическому классу можно поставить в соответствие определенный шаблон заполнения валентных гнезд. Это свойство было использовано в предлагаемой нами синтаксической модели предложения: в предикатную структуру введено поле, указывающее на семантический класс предиката. В нашей работе мы ориентировались на труды русского языковеда Л.М. Васильева [7]. В его «Системном семантическом словаре русского языка» предикатная лексика распределена на 12 основных семантических класса: 1) бытийные предикаты; 2) бытийно-пространственные предикаты (предикаты пространственной локализации); 3) предикаты отношения; 4) оценочные предикаты; 5) предикаты состояния; 6) количественные предикаты; 7) предикаты свойства; 8) предикаты поведения; 9) предикаты звучания; 10) предикаты движения; 11) акциональные предикаты; 12) акционально-процессуальные предикаты. Более того, в каждом из этих классов выделяют подклассы, т.е. предложенная классификация имеет иерархическую структуру.

С учетом вышесказанного синтаксическая модель предложения, которую мы предлагаем, описана следующей структурой:

$$PRED = <\{Obj_i\}\ i=1,...,7,\ sem>,$$

где $PRED$ – ядро структуры, предикат, sem – номер семантического класса, Obj_i – звезда, главное слово в ней субстантив, являющийся актантом.

Опишем этапы работы метода синтаксического анализа предложения русского языка, формирующего синтаксическую модель в виде структуры $PRED$.

Синтаксический анализ предложений

Анализ синтаксической структуры предложения должен выполняться на основе информации о словах, полученной на этапе графематического и морфологического анализа. При этом каждой словоформе предложения приписывается соответствующий набор (наборы – в случае морфологической омонимии) МИ. Таким образом, входными данными метода являются:

$$S = (s[1],..., s[i],..., s[N]),$$

где $s[i] = \{s[i][1],..., s[i][j],..., s[i][N]\}$ – вектор множеств интерпретаций словоформ, при этом каждое множество интерпретаций $s[i]$ является массивом пар (лемма, МИ).

Выходные данные с учетом синтаксической омонимии, в результате чего возможно получение нескольких вариантов синтаксического разбора, представляют собой множество пар вида (дерево зависимостей; предикатная модель).

Дерево зависимостей для предложения из N слов задается в матричном виде с помощью матрицы A, имеющей размерность NxN. Элементы матрицы, $a[i][j]$, представляют собой структуру, отражающую наличие и тип связи между словами $s[i]$ и $s[j]$, причем $s[i]$ – главное слово. Элемент $a[i][j]$ указывает на один из типов связи: атрибутивная (согласование), управление, примыкание, координация (отношение «подлежащее-сказуемое»). В свою очередь, связь «координация» описывается с помощью шаблона предикативного ядра простого предложения и имеет 17 типов (согласно количеству минимальных структурных схем простого предложения русского языка [8]). Подробно эти шаблоны и алгоритм их выделения описаны в работе [9].

Модуль синтаксического анализа осуществляет свою работу в несколько этапов:

1. Фрагментация – членение предложения по знакам пунктуации и союзам на сегменты, представляющие собой неразрывные синтаксические единства, и установление частичной иерархии на множестве этих единств. Подробно этот процесс изложен в [10]. Для работы на этом этапе используются словари шаблонов:

– обращений, вводных слов и конструкций, обособленных членов предложения;

– однородных членов предложения;

– употреблений союзов и союзных слов;

– для установления связанности пар сегментов.

2. Заполнение звезд: поиск пар потенциально связанных вариантов интерпретации словоформ, включая пару (грамматический предикат, грамматический субъект). Этот этап использует:

– правила выделения синтаксических связей пар слов;

– словарь шаблонов предикативного ядра простого предложения для выделения потенциальных синтаксических связей между главными членами предложения.

На выходе – наборы звезд: $<s[i], <s[j]>>$, где $s[i]$ – главное слово, $<s[j]>$ – множество зависимых слов.

3. Сокращение количества вариантов интерпретаций словоформ согласно критерию: для каждой словоформы хотя бы один вариант её интерпретации должен принадлежать либо множеству главных, либо множеству зависимых слов.

4. Заполнение актантной структуры найденного предиката. Заполняются семь валентных гнезд. Для чего используется семантический словарь предикатов, работа по созданию над которым ведется в настоящее время.

Опишем коротко состав словарной статьи. Поля статьи содержат данные о предикате следующего свойства:

1. Семантико-синтаксический класс.
2. Переходность (для глаголов).
3. Нуль- или не нуль-валентный.
4. Информация о заполнении валентных гнезд.

При заполнении валентных гнезд наряду с МИ актантов (как правило, являющимися субстантивами) указываются предлоги, которыми управляет предикат и которые управляют актантом. Следует обратить внимание, что актантом гнезд от 5-го до 7-го может быть наречие.

Например, для глагола «переправить»

1. Семантико-синтаксический класс 10.2.1.1 (глагол движения, обозначающий произвольное перемещение).
2. Переходный.
3. Не нуль-валентный.
4. Информация о заполнении валентных гнезд сведена в табл. 3.

Таблица 3 – Заполнение валентных гнезд для предиката «переправить»

Субъект	Объект	Адресат	Инструмент	Начальный локатив	Конечный локатив	Средний локатив
NULL 1	NULL 4	NULL 3	NULL 5	из 2 с 2 от 2	в 4 до 2 к 3	через 4

В табл. 3 NULL указывает на отсутствие предлога, цифра – на номер падежа субстантива, являющегося актантом, которым этот предлог управляет.

Выводы

Вопросы описания понятий входного текста, определение их свойств и отношений между ними должны решаться на уровне синтаксической модели, поскольку понятия и связи между ними проявляются в морфолого-синтаксических признаках и структурах и не зависят от смысла высказываний. Поэтому выбор используемой синтаксической модели крайне важен для проведения качественного семантического анализа.

Существующие способы представления синтаксических структур имеют определенные недостатки: деревья подчинения не учитывают связей между словосочетаниями и синтаксически целостными группами слов, системы НС игнорируют направленные связи и не позволяют описывать разрывные словосочетания. Кроме того, в этих представлениях члены предложения определяются на основе формальных признаков, а не по отношению к их семантическому содержанию. Поэтому ни одна из моделей не дает полного представления о синтаксической структуре предложения.

В данной работе предложена синтаксическая модель предложения в виде предикатной структуры, для формирования которой необходимо использовать лингвистические знания в виде семантического словаря предикатов, разработан метод синтаксического анализа, формирующий эту синтаксическую модель и опирающийся на словари шаблонов и набор правил выделения синтаксических связей пар слов.

Описанная в работе синтаксическая модель позволяет полностью выявлять как предикативные, так и синтагматические отношения, описывает не только аргументную структуру и количество актантов предиката, но также учитывает их семантическое содержание, используя семантическую классификацию предикатов.

Развитием данной работы может стать понимание текста, которое тесно связано с выявлением предикатных структур, характеризующих смысл предложений, а также – цепочек этих предикатных структур, которые опосредуют смысл текста. Полученные для множества текстов предметной области цепочки предикатных структур можно разбить на классы, которые характеризуют отдельные подобласти предметной области, и озаглавить названиями подобластей (подтем). Отнесение подцепочек цепочки предикатных структур, полученной для некоторого текста, к этим классам, и дальнейшая пометка их названиями соответствующих классов, и есть интерпретация текста, то есть, понимание.

Литература

1. Peter Jackson. Natural Language Processing for Online Applications / Peter Jackson, Isabelle Moulinier. – John Benjamins Publishing, 2002. — 237 p.
2. Ермаков А.Е. Выделение объектов в тексте на основе формальных описаний / А.Е. Ермаков, В.В. Плешко, В.А. Митюнин // Информационные технологии. - 2003. – N 12. – С. 1-6.
3. Дорохина Г.В. Модуль морфологического анализа без словаря слов русского языка / Г.В. Дорохина, В. Ю. Трунов, Е. В. Шилова // Искусственный интеллект. – 2010. – № 2. – С.32-36.
4. Ермаков А.Е. Компьютерная морфология в контексте анализа связного текста / А.Е. Ермаков, Плешко В.В. // Компьютерная лингвистика и интеллектуальные технологии : труды Международной конференции «Диалог'2004». – Москва : Наука, 2004. – С. 185-190.
5. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения / Гладкий А.В. – М : Наука, 1985. – 144 с.
6. Семантические типы предикатов / под ред. О.Н. Селиверстовой. – М. : Наука, 1982. – 365 с.
7. Васильев Л.М. Системный семантический словарь русского языка / Леонид Михайлович Васильев // Предикатная лексика. – Уфа : Изд-во «Восточный университет», 2000. − 200 с.
8. Современный русский язык : Учебник для филологических специальностей высших учебных заведений / В.А. Белошапкова, Е.А. Брызгунова, Е.А. Земская и др.; Под ред. Белошапковой ; [3-е изд, испр. и доп.] -– М. : Азбуковник, 1997. – 928 с.
9. Дорохина Г.В. Автоматическое выделение синтаксически связанных слов простого распространенного неосложненного предложения / Г.В. Дорохина, Д.С. Гнитько // Сучасна інформаційна Україна: інформатика, економіка, філософія : матеріали доповідей конференції, (12 - 13 травня 2011 року). – Донецьк, 2011. – Т. 1. – С. 34-38.
10. Сокирко А.В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) / Сокирко А.В. // Диссертация на соискание ученой степени кандидата технических наук. – МГПИИЯ.– М., 2001. – 108 с.

Literatura

1. Peter Jackson, Isabelle Moulinier. Natural Language Processing for Online Applications. — John Benjamins Publishing, 2002. — 237 p.
2. Ermakov A.E., Pleshko V.V., Mitjunin V.A. Vydelenie ob''ektov v tekste na osnove formal'nyh opisanij. // Informacionnye tehnologii. - 2003. – N 12. – S. 1-6.
3. Dorohina G. V. Modul' morfologicheskogo analiza bez slovarja slov russkogo jazyka / G. V. Dorohina, V. Ju. Trunov, E. V. Shilova // Iskusstvennyj intellekt. – №2. – 2010. – S. 32-36
4. Ermakov A.E., Pleshko V.V. Komp'juternaja morfologija v kontekste analiza svjaznogo teksta // Komp'juternaja lingvistika i intellektual'nye tehnologii: trudy Mezhdunarodnoj konferencii Dialog'2004. Moskva, Nauka, 2004 - S. 185-190.
5. Gladkij A.V. Sintaksicheskie struktury estestvennogo jazyka v avtomatizirovannyh sistemah obshhenija. - M: Nauka, - 1985, - 144 s.
6. Semanticheskie tipy predikatov / Pod red. O.N. Seliverstovoj. – M.: Nauka, 1982. – 365 s.
7. Vasil'ev L.M. Sistemnyj semanticheskij slovar' russkogo jazyka / Leonid Mihajlovich Vasil'ev // Predikatnaja leksika. – Ufa: Izd – vo «Vostochnyj universitet», 2000. − 200 s.
8. Sovremennyj russkij jazyk: Uchebnik dlja filologicheskih special'nostej vysshih uchebnyh zavedenij / V.A. Beloshapkova, E.A. Bryzgunova, E.A. Zemskaja i dr.; Pod red. Beloshapkovoj – 3-e izde, ispr. i dop. - M.: Azbukovnik, 1997 - 928 s.
9. Dorohina G. V. Avtomaticheskoe vydelenie sintaksicheski svjazannyh slov prostogo rasprostranennogo neoslozhnennogo predlozhenija / G.V. Dorohina, D. S. Gnit'ko // «Suchasna informacijna Ukraïna: informatika, ekonomika, filosofija»: materiali dopovidej konferenciï, 12 - 13 travnja 2011 roku, Donec'k, 2011. T. 1. – S. 34-38.
10. Sokirko A.V. Semanticheskie slovari v avtomaticheskoj obrabotke teksta (po materialam sistemy DIALING) // Dissertacija na soiskanie uchenoj stepeni kandidata tehnicheskih nauk. – MGPIIJa. — M., 2001. — 108 s.

RESUME

T.V. Yermolenko, A.S. Gajdamaka

Syntactic Model of Russian Sentence Based on Predicate-Argument Structure

For today the problems of automatic text processing (ATP), that relating to unstructured data analysis for large arrays of documents processing (information retrieval, classification, cluster analysis, detection hidden patterns of relationship, etc.), effectively are solved by a wide class of statistical methods. As a result, a linguistic component of the algorithms stood aside pro tempore. Further complicating of mathematic without serious linguistics does not allow much to improve the quality of such systems. Using natural language in practical areas is impossible without equipment of ATP-system with wide and deep (in terms of coverage of different language levels) descriptions and models.

The article is devoted to the problems of development of a syntactic parser, it is one of the most important modules of ATP-systems, the semantic text processing is impossible without such parser.

The authors propose a syntactic model of sentence presentation that unites syntactic and semantic components. This model is a predicate-argument structure and is formed on the basis of the two relation types: predictive and syntagmatic.

Subject and predicate are marked out on the basis of minimum structure scheme of sentence vocabulary to form a predicate-argument structure of sentence. The valence slots are completed using a linguistic knowledge in the form of a semantic dictionary of predicates, actants act as a semantic cases, the number of valence determines its type, semantics and morphological characteristics.

Method of parsing is developed, it forms the predicate-argument structure of sentence based on templates dictionary and ruleset for isolating syntax relations between pairs of words.

Статья поступила в редакцию 05.07.2012.